AD629518

Working Paper No. 89

On The Asymptotic Behavior of k-means

by

James B. MacQueen

November 1965

Western Management Science Institute
University of California • Los Angeles

University of California

Los Angeles

Western Management Science Institute

Working Paper No. 89

On The Asymptotic Behavior of k-means

by

James B. MacQueen

November 1965

This work was supported by the Western Management Science Institute under a grant from the Ford Foundation, and by the Office of Naval Research under Contract No. 233(75), Task No. 047-041.

ON THE ASYMPTOTIC BEHAVIOR

OF K-MEANS

J. MacQueen

University of California, Los Angeles

1. Introduction. Let $z_1, z_2, \ldots$ be a random sequence of points (vectors) in E_N, each point being selected independently of the preceding ones using a fixed probability measure p. Thus $P[z_1 \epsilon A] = p(A)$ and $P[z_{n+1} \epsilon A | z_1, z_2, \ldots, z$ $= p(A)$, $n=1,2,\ldots$, for A any measurable set in E_N. Relative to a given k-tuple $x = (x_1, x_2, \ldots, x_k)$, $x_i \epsilon E_N$, $i = 1,2,\ldots k$, we define a minimum distance partitior $S(x) = \{S_1(x), S_2(x), \ldots, S_k(x)\}$ of E_N, by $S_1(x) = T_1(x)$, $S_2(x) = T_2(x)S_1'(x), \ldots,$ $S_k(x) = T_k(x)\, S_1'(x)S_2'(x)\ldots S_{k-1}'(x)$, where $T_i(x) = \{\xi : \xi \,\epsilon\, E_N,$ $|\xi - x_i| \leq |\xi - x_j|$, $j = 1,2,\ldots,k\}$. The set $S_i(x)$ contains the points in E_N nearest to x_i, with tied points being assigned arbitrarily to the set of lower index. Note that with this convention concerning tied points, if $x_i = x_j$ and $i < j$ then $S_j(x) = \emptyset$. Sample k-means $x^n = (x_1^n, x_2^n, \ldots, x_k^n)$, $x_i^n \,\epsilon\, E_N$, $i = 1,\ldots,k$, with associated integer weights $(w_1^n, w_2^n, \ldots w_k^n)$, are now defined as follows: $x_i^1 = z_i, w_i^1 = 1$, $i = 1,2,\ldots,k$, and for $n = 1,2,\ldots$ if $z_{k+n} \epsilon S_i^n$, $x_i^{n+1} = (x_i^n w_i^n + z_{n+k})/(w_i^n + 1)$, $w_i^{n+1} = w_i^n + 1$, and $x_j^{n+1} = x_j^n$, $w_j^{n+1} = w_j^n$ for $j \neq i$, where $S^n = \{S_1^n, S_2^n, \ldots, S_k^n\}$ is the minimum distance partition relative to x^n.

We investigate the asymptotic behavior of the k-means, making the special assumptions, (i), p is absolutely continuous with respect to Lebesgue measure on E_N, and (ii), $p(R) = 1$ for a closed and bounded convex set $R \subset E_N$, and $p(A) > 0$ for every open set $A \subset R$. For a given k-tuple $x = (x_1, x_2, \ldots x_k)$ -- such an entity being referred to hereafter as a k-point -- let

$$W(x) = \Sigma_{i=1}^{k} \int_{S_i} |z - x_i|^2 dp(z) ,$$

$$V(x) = \Sigma_{i=1}^{k} \int_{S_i} |z - u_i(x)|^2 dp(z) ,$$

---

1 This work was supported by the Western Management Science Institute under a grant from the Ford Foundation, and by the Office of Naval Research under Contract No. 233(75), Task No. 047-041.

where $S=\{S_1,S_2,\ldots S_k\}$ is the minimum distance partition relative to x, and $u_i(x) = \int_{S_i} z dp(z)/p(S_i)$ or $u_i(x) = x_i$ according as $p(S_i) > 0$ or $p(S_i) = 0$. If $x_i = u_i(x)$, $i = 1,2,\ldots,k$, we say the k-point x is unbiased.

The principle result is

Theorem 1. *The sequence of random variables* $W(x^1)$, $W(x^2)$, ... *converges a.s. and* $W_\infty = \lim_{n\to\infty} W(x^n)$ *is a.s. equal to* $V(x)$ *for some* x *in the class of k-points* $x=(x_1,x_2,\ldots x_k)$ *which are unbiased, and have the property that* $x_i \neq x_j$ *if* $i \neq j$.

In lieu of a satisfactory strong law of large numbers for k-means, we obtain

Theorem 2. $\Sigma_{n=1}^m (\Sigma_{i=1}^k p_i^n |x_i^n - u_i^n|)/m \underset{a.s.}{\rightarrow} 0$ *as* $m \to \infty$ *where* $u_i^n = u_i(x^n)$ *and* $p_i^n = p(S_i(x^n))$.

Potential applications of the k-means concept, which will be discussed in detail elsewhere, occur in certain taxomony problems, in connection with coding and pattern recognition problems, in the description of categorizing behavior, and in connection with the problem of locating partitions with minimum average variance [5] (See Box [1] and Ward [6] for related results.).

2. Proofs. The system of k-points forms a complete metric space if the distance $\rho(x,y)$ between the k-points $x = (x_1,x_2,\ldots x_k)$ and $y = (y_1,y_2,\ldots y_k)$, is defined by $\rho(x,y) = \Sigma_{i=1}^k d(x_i,y_i)$, where $d(a,b)$ is the Euclidian distance between a and b. We designate this space by M and interpret continuity, limits, convergence, neighborhoods, etc., in the usual way with respect to the metric topology of M. Of course, every bounded sequence of k-points contains a convergent subsequence.

Certain difficulties encountered in the proof of Theorem 1 are caused by the possibility of the limit of a convergent sequence of k-points having some of its constituent points equal to each other. With the end in view of circumventing these difficulties, suppose that for a given k-point $x = (x_1, x_2, \ldots, x_k)$, $x_i \epsilon R$, $i=1,2,\ldots,k$, we have $x_i = x_j$ for a certain pair $i, j, i<j$, and $x_i = x_j \neq x_m$ for $m \neq i, \neq j$. The points x_i and x_j being distinct in this way, and considering assumption (ii) we necessarily have $p(S_i(x)) > 0$, for $S_i(x)$ certainly contains an open sub-set of R. The convention concerning tied points means $p(S_j(x)) = 0$. Now if $\{y^n\} = \{(y_1^n, y_2^n, \ldots, y_k^n)\}$ is a sequence of k-points satisfying $y_i^n \epsilon R$, and $y_i^n \neq y_j^n$ if $i \neq j$, $n=1,2,\ldots,$ and the sequence y^n approached x, then y_i^n and y_j^n approach $x_i = x_j$, and hence each other; they also approach the boundaries of $S_i(y^n)$ and $S_j(y^n)$ in the vicinity of x_i. The conditional means $u_i(y^n)$ and $u_j(y^n)$, however, must remain in the interior of the sets $S_i(y^n)$ and $S_j(y^n)$ respectively, and thus tend to become separated from the corresponding points y_i^n and y_j^n. In fact, for each sufficiently large n, the distance of $u_i(y^n)$ from the boundary of $S_i(y^n)$ <u>or</u> the distance of $u_j(y^n)$ from the boundary of $S_j(y^n)$, will exceed a certain positive number. For as n tends to infinity, $p(S_i(y^n)) + p(S_j(y^n))$ will approach $p(S_i(x)) > 0$ -- a simple continuity argument based on the absolute continuity of p will establish this -- and for each sufficiently large n, at least one of the probabilities $p(S_i(y^n))$ or $p(S_j(y^n))$ will be positive by a definite amount, say δ. But in view of the boundedness of R, a convex set of p measure at least $\delta > 0$ cannot have its conditional mean arbitrarily near its boundary. This line of reasoning, which extends

immediately to the case where some three or more members of $(x_1,x_2,\ldots x_k)$ are equal, gives us

Lemma 1. Let $x = (x_1,x_2,\ldots x_k)$ be the limit of a convergent sequence of k-points $\{y^n\} = \{(y_1^n,y_2^n,\ldots y_k^n)\}$ satisfying $y_i^n \in R$, $y_i^n \neq y_j^n$ if $i \neq j$, $n=1,2,\ldots$. If $x_i=x_j$ for some $i\neq j$ then $\lim\inf_n \Sigma_{i=1}^k p(S_i(y^n))|y_i^n - u_i(y^n)| > 0$. Hence, if $\lim_{n\to\infty} \Sigma_{i=1}^k p(S_i(y^n))|y_i^n - u_i(y^n)| = 0$, each member of the k-tuple $(x_1,x_2,\ldots x_k)$ is distinct from the others.

We remark that if each member of the k-tuple $x=(x_1,x_2,\ldots x_k)$ is distinct from the others, then $\pi(y) = (p(S_1(y)), p(S_2(y)),\ldots p(S_k(y))$, regarded as a mapping of M onto E_k, is continuous at x-- this follows directly from the absolute continuity of p. Similarly $u(y) = (u_1(y), u_2(y),\ldots u_k(y))$ regarded as a mapping from M onto M is continuous at x -- because of the absolute continuity of p and the boundness of R (finiteness of $\int z dp(z)$ would do.) Putting this remark together with Lemma 1, we get

Lemma 2. Let $x = (x_1,x_2,\ldots x_k)$ be the limit of a convergent sequence of k-points $\{y^n\} = \{(y_1^n, y_2^n,\ldots y_k^n)\}$ satisfying $y_i^n \in R$, $y_i^n \neq y_j^n$ if $i \neq j$. $n=1,2,\ldots$. If $\lim_{n\to\infty}\Sigma_{i=1}^k p(S_i(y^n))|y_i^n - u_i(y^n)| = 0$ then $\Sigma_{i=1}^k p(S_i(x))|x_i - u_i(x^n)| = 0$ and each point x_i in the k-tuple $(x_1,x_2,\ldots x_k)$ is distinct from the others.

Lemma 1 and 2 above are primarily technical in nature. The heart of the proofs of theorem 1 and 2 is the following application of Martingale theory:

Lemma 3. Let $t_1, t_2,\ldots,$ and $\xi_1,\xi_2,\ldots$ be given sequences of random variables, and for each $n=1,2,\ldots,$ let t_n be measurable with respect to β_n where $\beta_1\subset \beta_2\subset$ is a monotone increasing sequence of σ- fields (belonging to the underlying probability space). Suppose each of the following conditions holds a.s.: (i) $|t_n| \leq K < \infty$, (ii) $\xi_n \geq 0$, $\Sigma \xi_n < \infty$, (iii) $E(t_{n+1}|\beta_1,\beta_2,\ldots\beta_n) \leq t_n + \xi_n$. Then the sequences of random variables $t_1,t_2,\ldots$ and $s_0,s_1,s_2,\ldots,$

where $s_0 = 0$ *and* $s_n = \Sigma_{i=1}^n (t_i - E(t_{i+1}|\beta_1,\beta_2,\ldots\beta_i)$, $n = 1,2,\ldots,$ *both converge a.s.*

Proof. Let $y_n = t_n + s_{n-1}$ so that the y_n form a Martingale sequence. Let c be a positive number and consider the sequence $\{\tilde{y}_n\}$ obtained by stopping y_n (see [2], p. 300) at the first n for which $y_n \leq -c$. From (iii) we see that $y_n \geq -\Sigma_{i=1}^{n-1} \xi_i - K$ and since $y_n - y_{n-1} \geq 2K$, we have $\tilde{y}_n \geq \max(-\Sigma_{i=1}^{n-1} \xi_i - K, -(c+2K))$. The sequence $\{\tilde{y}\}$ is a Martingale, so that $E\tilde{y}_n = E\tilde{y}_1$, $n=1,2,\ldots,$ and being bounded from below with $E|\tilde{y}_1| \leq K$, certainly $\sup_n E|\tilde{y}_n| < \infty$. The Martingale Theorem [2, p. 319] shows $\tilde{y}_n$ converges a.s. But $y_n = \tilde{y}_n$ on the set A_c where $-\Sigma_{i=1}^{\infty} \xi_i > -c - K$, $i = 1,2,\ldots,$ and (ii) implies $P[A_c] \to 1$ as $c \to \infty$. Thus $\{y_n\}$ converge a.s. This means $s_n = y_{n+1} - t_{n+1}$ is a.s. bounded. Using (iii) we can write $-s_n = \Sigma_{i=1}^n \xi_i - \Sigma_{i=1}^n \Delta_i$ where $\Delta_i \geq 0$. But since s_n and $\Sigma_1^n \xi_i$ are a.s. bounded, $\Sigma \Delta_i$ converges a.s., s_n converges a.s., and finally, so does t_n. This completes the proof.

Turning now to the proof of Theorem 1, let ω_n stand for the sequence $z_1, z_2, \ldots z_{n-1+k}$, and let A_i^n be the event $[z_{n+k} \in S_i^n]$. Since S^{n+1} is the minimum distance partition relative to x^{n+1}, we have

$$(1) \qquad E[W(x^{n+1})|\omega_n] = E[\Sigma_{i=1}^k \int_{S_i^{n+1}} |z - x_i^{n+1}|^2 dp(z)|\omega_n]$$

$$\leq E[\Sigma_{i=1}^k \int_{S_i^n} |z - x_i^{n+1}|^2 dp(z)|\omega_n]$$

$$= \Sigma_{j=1}^k E[\Sigma_{i=1}^k \int_{S_i^n} |z - x_i^{n+1}|^2 dp(z)|A^r, \omega_n]p_j^n .$$

If $z_{n+k} \in S_j^n$, $x_i^{n+1} = x_i^n$ for $i \neq j$. Thus we obtain

$$(2) \qquad E[W(x^{n+1})|\omega_n] \leq W(x^n) - \Sigma_{j=1}^{k}(\int_{S_j^n}|z-x_j^n|^2 dp(z))p_j^n$$

$$+ \Sigma_{j=1}^{k} E[\int_{S_j^n}|z - x_j^{n+1}|^2 dp(z)|A_j^n, \omega_n]p_j^n .$$

Several applications of the relation $\int_A|z-x|^2dp(z) = \int_A|z-u|^2dp(z) + p(A)|x-u|^2$, where $\int_A(u-z)\,dp(z) = 0$, enables us to write the last term in (2) as

$$\Sigma_{j=1}^{k}[\int_{S_j^n}|z-x_j^n|^2\,dp(z)p_j^n - (p_j^n)^2|x_j^n - u_j^n|^2$$

$$+ (p_j^n)^2|x_j^u - u_j^n|^2(w_j^n/w_j^n+1))^2 + \int_{S_j^n}|z - u_j^n|^2dp(z)p_j^n/(w_j^n + 1)^2].$$

Combining this with (2), we get

$$(3) \qquad E(W(x^{n+1}) \mid \omega_n] \leq W(x^n) - \Sigma_{j=1}^{k}|x_j^n - u_j^n|^2(p_j^n)^2(2w_j^n + 1)/(w_j^n + 1)^2$$

$$+ \Sigma_{j=1}^{k} \sigma_{n,j}^2(p_j^n)^2/(w_j^n + 1)^2 ,$$

where $\sigma_{n,j}^2 = \int_{S_j^n}|z - u_j^n|^2dp(z)/p_j^n$.

Since we are assuming $p(R) = 1$, certainly $W(x^n)$ is a.s. bounded, as is $\sigma_{n,j}^2$. We now show that

$$(4) \qquad \Sigma_n(p_j^n)^2/(w_j^n + 1)^2$$

converges a.s. for each $j=1,2,\ldots k$, thereby showing that $\Sigma_n(\Sigma_{j=1}^{k}[\sigma_{n,j}^2(p_j^n)^2/(w_j^n + 1)^2]$ converges a.s. Then Lemma 3 can be applied with $t_n = W(x^n)$ and $\xi_n = \Sigma_{j=1}^{k}\sigma_{n,j}^2(p_j^n)^2/(w_j^n + 1)^2$.

It suffices to prove that

$$(5) \qquad \Sigma_{n\geq 2}(p_j^n)^2/[(\beta + 1 + w_j^n)(\beta + 1 + w_j^{n+1})]$$

converges a.s. for any positive number β; also, this is convenient, for $E(I_j^n|\omega_n) = p_j^n$ where I_j^n is the characteristic function of the event $[z_{n+k} \varepsilon S_j^n]$, and on noting that $w_j^{n+1} = 1 + \Sigma_{i=1}^{n} I_j^i$, a direct application

of Theorem 1, p. 274, in [3], says that for any positive numbers α and β, $P[\beta+1+w_j^{n+1} \geq 1 + \Sigma_{i=1}^n p_j^i - \alpha\Sigma_{i=1}^n v_j^i$ <u>for all</u> $n=1,2,\ldots] > 1 - (1+\alpha\beta)^{-1}$, where $v_j^i = p_j^i - (p_j^i)^2$ is the conditional variance of I_j^i given ω_i. We take $\alpha=1$, and thus with probability at least $1 - (1+\beta)^{-1}$ the series (5) is dominated by

$$\Sigma_{n\geq 2} (r_j^n)^2/[(1 + \Sigma_{i=1}^{n-1} (p_j^i)^2)(1+\Sigma_{i=1}^n(p_j^i)^2)]$$
$$= \Sigma_{n\geq 2}[1/(1+\Sigma_{i=1}^{n-1}(p_j^i)^2) - 1/(1+\Sigma_{i=1}^n(p_j^i)^2)] ,$$

which clearly converges.

The choice of β being arbitrary, we have shown that (4) converges a.s. Application of Lemma 3 as indicated above proves $W(x^n)$ converges a.s.

To identify the limit W_∞, note that with t_n and ξ_n taken as above, Lemma 3 entails a.s. convergence of $\Sigma_n[W(x^n) - E[W(x^{n+1})|\omega_n]]$, and hence (3) implies a.s. convergence of

(6) $\quad \Sigma_n(\Sigma_{j=1}^k|x^n_j - u_j^n|^2(p_j^n)^2 (2w_j^n + 1)/(w_j^n + 1)^2)$.

Since (6) dominates $\Sigma_n(\Sigma_{j=1}^k p_j^n|x_j^n - u_j^n)/kn$, the latter converges a.s., and a little consideration makes it clear that $\Sigma_{j=1}^k p_j^n|x_j^n - u_j^n| = \Sigma_{j=1}^k p(S_j(x^n))|x_j^n - u_j(x^n)|$ converges to zero on a sub-sequence $\{x^{n_s}\}$ and that this sub-sequence has itself a convergent sub-sequence, say $\{x^{n_t}\}$. Let $x = (x_1,x_2,\ldots x_k) = \lim_{t\to\infty}x^{n_t}$. Since $W(x) = V(x) + \Sigma_{j=1}^k p(S_j(x))|x_j-u(x)|^2$ and in particular $W(x^n) = V(x^n) + \Sigma_{j=1}^k p(S_j(x^n))|x_j^n - u(x_j^n)|^2$, we have only to show (a), $\lim_{t\to\infty} W(x^{n_t}) = W_\infty = W(x)$, and (b), $\lim_{t\to\infty}\Sigma_{j=1}^k p(S_j(x^{n_t}))|x_j^{n_t}-u(x_j^{n_t})|^2 = 0 = \Sigma_{j=1}^k p(S_j(x))|x_j - u_j(x)|^2$. Then $W(x) = V(x)$ and x is a.s. unbiased. (Obviously $\Sigma_{i=1}^k p_i|a_i| = 0$ if and only if $\Sigma_{i=1}^k p_i|a_i|^2 = 0$, where $p_i \geq 0$.)

We show that (a) is true by establishing the continuity of $W(x)$. We have

$$W(x) \leq \Sigma_{j=1}^{k} \int_{S_j(y)} |z-x_j|^2 dp(z)$$

$$\leq \Sigma_{j=1}^{k} \int_{S_j(y)} |z-y_j|^2 + \Sigma_{j=1}^{k} [p(S_j(y))|x_j - y_j|^2 +$$

$$+ 2|x_j - y_j| \int_{S_j(y)} |z - x_j| dp(z)],$$

with the last inequality following easily from the triangle inequality. Thus $W(x) \leq W(y) + o(\rho(x,y))$, and similarly $W(y) \leq W(x) + o(\rho(x,y))$.

To establish (b), Lemma 2 can be applied with $\{y^n\}$ and $\{x^{n_t}\}$ identified, for a.s. $x_i^n \neq x_j^n$ for $i \neq j$, $n=1,2,\ldots$. It remains to remark that Lemma 2 also implies a.s. $x_i \neq x_j$ for $i \neq j$. The proof of Theorem 1 is complete.

Theorem 2 follows from the a.s. convergence of $\Sigma_n(\Sigma_{i=1}^{k} p_i^n |x_i^n - u_i^n|)/nk$ upon applying an elementary result, (c.f. Theorem C, p. 203 in [4]) which says that if $\Sigma\, a_n/n$ converges, $\Sigma_{i=1}^{n} a_i/n \to 0$.

3. <u>Remarks</u>. In a number of cases covered by Theorem 1, all the unbiased k-points have the same value of W. In this situation, Theorem 1 implies $\Sigma_{i=1}^{k} p_i^n |x_i^n - u_i^n|$ converges a.s. to zero. An example is provided by the uniform distribution over a disk in E_2. If $k = 2$, the unbiased k-points (x_1, x_2) with $x_1 \neq x_2$ consist of the family of points x_1 and x_2 opposite one another on a diameter, and at a certain fixed distance from the center of the disk. (There is one unbiased k-point with $x_1 = x_2$, both x_1 and x_2 being at the center of the disk in this case.) The k-means thus converge to some such relative position, but Theorem 1 does not quite permit us to eliminate the interesting possibility that the two means oscillate slowly but indefinitely around the center.

Theorem 1 provides for a.s. convergence of $\Sigma_{i=1}^{k} p_i^n |x_i^n - u_i^n|$ to zero in a slightly broader class of situations: This is where the unbiased k-points $x = (x_1, x_2, \ldots x_k)$ with $x_i \neq x_j$ for $i \neq j$, are all stable in the sense that for each such x, $W(y) \geq W(x)$ (and hence $V(y) \geq V(x)$) for all y in a neighborhood of x. In this case, each such x falls in one of finitely many equivalence classes such that W is constant on each class. This is illustrated by the above example, where there is only a single equivalence class. If each of the equivalence classes contains only a single point, Theorem 1 implies a.s. convergence of x^n to one of those points.

There are unbiased k-points which are not stable. Take a distribution on E_2 which has sharp peaks of probability at each corner of a square, and is symetric about both diagonals. With k=2, the two constituent points can be symetrically located on a diagonal so that the boundary of the associated minimum distance partition coincides with the other diagonal. With some adjustment, such a k-point can be made to be unbiased, and if the probability is sufficiently concentrated at the corners of the square, any small movement of the two points off the diagonal in opposite directions, results in a decrease in W(x). It seems likely that the k-means cannot converge to such a configuration.

REFERENCES

1. Cox, D.R., (1957) Note on grouping. J. Amer. Stat. Assoc. 52(2), 543-547 .

2. Doob, J. L., (1953) Stochastic processes. John Wiley & Sons, New York.

3. Dubins, L. E. and Savage, L. J. (1965) A Techebycheff-like inequality for stochastic processes. Proceedings Nat. Ac. Scien. 53(2) 274-275.

4. Halmos, Paul R., (1950) Measure theory. Van Nostrand, New York.

5. MacQueen, J. (1965) On convergence of k-means and partitions with minimum average variance. (Abstract of paper presented at the Western regional meetings of the Institute of Mathematical Statistics, Berkeley, June 19, 1965.) Ann. Math Stat. 36(3) 1084.

6. Ward, Joe., (1963) Hierarchical grouping to optimize an objective function. J. Amer. Stat. Assoc. 58, 301, 236-244.

BLANK PAGE

Security Classification

DOCUMENT CONTROL DATA - R&D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1 ORIGINATING ACTIVITY (Corporate author) Western Management Science Institute University of California Los Angeles, California 90024	2a REPORT SECURITY CLASSIFICATION Unclassified 2b GROUP

3 REPORT TITLE

On The Asymptotic Behavior of K-means

4 DESCRIPTIVE NOTES (Type of report and inclusive dates)

Working Paper

5 AUTHOR(S) (Last name, first name, initial)

MacQueen, James B.

6 REPORT DATE	7a TOTAL NO. OF PAGES	7b NO. OF REFS
November 1965	10	6

8a CONTRACT OR GRANT NO. 233(75) b. PROJECT NO. 047-041 c d	9a ORIGINATOR'S REPORT NUMBER(S) Working Paper No. 89 9b OTHER REPORT NO(S) (Any other numbers that may be assigned this report)

10 AVAILABILITY/LIMITATION NOTICES

Available upon request through: Western Management Science Institute
University of California
Los Angeles, California

11 SUPPLEMENTARY NOTES	12 SPONSORING MILITARY ACTIVITY

13 ABSTRACT

For a sample sequence $y_1, y_2, \ldots$ representing independent observations on an N-dimensial r.v.y, define sample k-means $x^n = \{x_1^n, x_2^n, \ldots, x_k^n\}$ with weights $w^n = \{w_1^n, w_2^n, \ldots w_k^n\}$ as follows: $x_i^1 = y_i$, $w_i^1 = 1$, $i = 1,2,\ldots k$, x^{n+1}, w^{n+1} are formed from x^n, w^n by the rule that if y_{k+n+1} is nearest to x_i^n, then $x_i^{n+1} = (x_i^n w_i^n + y_{k+n+1})/(w_i^n + 1)$, $w_i^{n+1} = w_i^n + 1$, and $x_j^{n+1} = x_j^n$, $w_j^{n+1} = w_j^n$, $j \neq i$.

The asymptotic behavior of the k-means is studied and it is shown that $\Sigma_{i=1}^k \int_{S_i^n} |z - x_i^n|^2 \, dp(z)$ converges a.s., where S_i^n is the region in E_N nearer to x_i^n than x_j^n, $j \neq i$, and p is the common probility measure of the y_i.

Applications of the k-means concept occur in statistical analysis of N-dimensional data, in coding problems, and in the discription of human judgement.

DD FORM 1 JAN 64 1473 0101-807-6800

Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	
Classification Grouping Clustering Prediction Relevant Non-Linear Multivariate Partitioning						

INSTRUCTIONS

1. ORIGINATING ACTIVITY: Enter the name and address of the contractor, subcontractor, grantee, Department of Defense activity or other organization *(corporate author)* issuing the report.

2a. REPORT SECURITY CLASSIFICATION: Enter the overall security classification of the report. Indicate whether "Restricted Data" is included. Marking is to be in accordance with appropriate security regulations.

2b. GROUP: Automatic downgrading is specified in DoD Directive 5200.10 and Armed Forces Industrial Manual. Enter the group number. Also, when applicable, show that optional markings have been used for Group 3 and Group 4 as authorized.

3. REPORT TITLE: Enter the complete report title in all capital letters. Titles in all cases should be unclassified. If a meaningful title cannot be selected without classification, show title classification in all capitals in parenthesis immediately following the title.

4. DESCRIPTIVE NOTES: If appropriate, enter the type of report, e.g., interim, progress, summary, annual, or final. Give the inclusive dates when a specific reporting period is covered.

5. AUTHOR(S): Enter the name(s) of author(s) as shown on or in the report. Enter last name, first name, middle initial. If military, show rank and branch of service. The name of the principal author is an absolute minimum requirement.

6. REPORT DATE: Enter the date of the report as day, month, year; or month, year. If more than one date appears on the report, use date of publication.

7a. TOTAL NUMBER OF PAGES: The total page count should follow normal pagination procedures, i.e., enter the number of pages containing information.

7b. NUMBER OF REFERENCES: Enter the total number of references cited in the report.

8a. CONTRACT OR GRANT NUMBER: If appropriate, enter the applicable number of the contract or grant under which the report was written.

8b, 8c, & 8d. PROJECT NUMBER: Enter the appropriate military department identification, such as project number, subproject number, system numbers, task number, etc.

9a. ORIGINATOR'S REPORT NUMBER(S): Enter the official report number by which the document will be identified and controlled by the originating activity. This number must be unique to this report.

9b. OTHER REPORT NUMBER(S): If the report has been assigned any other report numbers *(either by the originator or by the sponsor)*, also enter this number(s).

10. AVAILABILITY/LIMITATION NOTICES: Enter any limitations on further dissemination of the report, other than those imposed by security classification, using standard statemen such as:

(1) "Qualified requesters may obtain copies of this report from DDC."

(2) "Foreign announcement and dissemination of this report by DDC is not authorized."

(3) "U. S. Government agencies may obtain copies of this report directly from DDC. Other qualified DD users shall request through

________________________."

(4) "U. S. military agencies may obtain copies of this report directly from DDC. Other qualified users shall request through

________________________."

(5) "All distribution of this report is controlled. Qualified DDC users shall request through

________________________."

If the report has been furnished to the Office of Techni Services, Department of Commerce, for sale to the public, i cate this fact and enter the price, if known.

11. SUPPLEMENTARY NOTES: Use for additional expla tory notes.

12. SPONSORING MILITARY ACTIVITY: Enter the name the departmental project office or laboratory sponsoring *(pa ing for)* the research and development. Include address.

13. ABSTRACT: Enter an abstract giving a brief and fact summary of the document indicative of the report, even tho it may also appear elsewhere in the body of the technical r port. If additional space is required, a continuation sheet be attached.

It is highly desirable that the abstract of classified re be unclassified. Each paragraph of the abstract shall end an indication of the military security classification of the i formation in the paragraph, represented as *(TS), (S), (C), o*

There is no limitation on the length of the abstract. H ever, the suggested length is from 150 to 225 words.

14. KEY WORDS: Key words are technically meaningful t or short phrases that characterize a report and may be used index entries for cataloging the report. Key words must be selected so that no security classification is required. Ide fiers, such as equipment model designation, trade name, mi project code name, geographic location, may be used as ke words but will be followed by an indication of technical co text. The assignment of links, roles, and weights is optio